

PATENT  
8001-1176

**IN THE U.S. PATENT AND TRADEMARK OFFICE**

In re application of: Atsushi KUWATA  
Conf.:  
Appl. No.: NEW NON-PROVISIONAL  
Group:  
Filed: November 25, 2003  
Examiner:  
Title: DISK ARRAY APPARATUS AND DATA WRITING  
METHOD USED IN THE DISK ARRAY APPARATUS

CLAIM TO PRIORITY

Assistant Commissioner for Patents  
P.O. Box 1450  
Alexandria, VA 22313-1450

November 25, 2003

Sir:

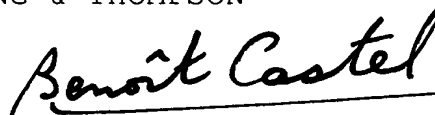
Applicant(s) herewith claim(s) the benefit of the  
priority filing date of the following application(s) for the  
above-entitled U.S. application under the provisions of 35  
U.S.C. § 119 and 37 C.F.R. § 1.55:

<u>Country</u>	<u>Application No.</u>	<u>Filed</u>
JAPAN	2003-001314	January 7, 2003

Certified copy(ies) of the above-noted application(s)  
is(are) attached hereto.

Respectfully submitted,

YOUNG & THOMPSON



Benoit Castel, Reg. No. 35,041

745 South 23<sup>rd</sup> Street  
Arlington, VA 22202  
Telephone (703) 521-2297

BC/ia

Attachment(s): 1 Certified Copy(ies)

日 本 国 特 許 庁  
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出 願 年 月 日                      2 0 0 3 年    1 月    7 日  
Date of Application:

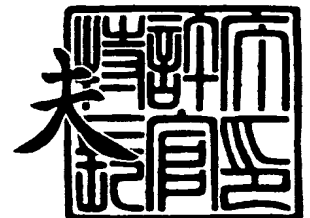
出 願 番 号                      特 願 2 0 0 3 - 0 0 1 3 1 4  
Application Number:  
[ST. 10/C] :                      [ J P 2 0 0 3 - 0 0 1 3 1 4 ]

出      願      人                      日 本 電 気 株 式 有 限 公 司  
Applicant(s):

2 0 0 3 年 1 0 月    2 日

特許庁長官  
Commissioner,  
Japan Patent Office

今 井 康 夫



出証番号    出証特 2 0 0 3 - 3 0 8 1 3 1 0

【書類名】 特許願

【整理番号】 67000110

【提出日】 平成15年 1月 7日

【あて先】 特許庁長官 殿

【国際特許分類】 G06F 03/06

【発明の名称】 ディスクアレイ装置及びディスクアレイ装置におけるデータ書き込み方法

【請求項の数】 7

【発明者】

    【住所又は居所】 東京都港区芝五丁目 7 番 1 号 日本電気株式会社内

    【氏名】 桑田 篤史

【特許出願人】

    【識別番号】 000004237

    【氏名又は名称】 日本電気株式会社

【代理人】

    【識別番号】 100079164

    【弁理士】

    【氏名又は名称】 高橋 勇

    【電話番号】 03-3862-6520

【手数料の表示】

    【予納台帳番号】 013505

    【納付金額】 21,000円

【提出物件の目録】

    【物件名】 明細書 1

    【物件名】 図面 1

    【物件名】 要約書 1

    【包括委任状番号】 9003064

【プルーフの要否】 要

【書類名】 明細書

【発明の名称】 ディスクアレイ装置及びディスクアレイ装置におけるデータ書き込み方法

【特許請求の範囲】

【請求項 1】 上位ホストからの指令により複数のディスクに対してデータを読み書き制御する制御部と、前記ディスクに対して読み書きするデータを一時的に記憶するキャッシュメモリとを備え、

前記制御部が、前記キャッシュメモリ上において、前記上位ホストにて用いられる論理アドレスに関連付けたデータを物理アドレスに関連付けて前記ディスクに対して読み書き制御を行うディスクアレイ装置において、

前記制御部が、前記ディスクに対して読み書き制御を行う際に、前記キャッシュメモリ上の前記物理アドレスに関連付けられたデータを当該物理アドレスに対応する前記ディスク上のデータに対して優先して処理する、ことを特徴とするディスクアレイ装置。

【請求項 2】 前記制御部が、前記ディスクに対してデータの書き込み処理を行う前に、当該ディスクに書き込むデータを物理アドレスに関連付けて前記キャッシュメモリに格納する、ことを特徴とする請求項 1 記載のディスクアレイ装置。

【請求項 3】 前記制御部が、前記ディスクにデータの書き込み処理を行うと共に当該書き込みが完了したことを確認した後に、前記キャッシュメモリ上で前記物理アドレスに関連付けられた書き込みデータを当該物理アドレスに関連付けられた状態から解除する、ことを特徴とする請求項 2 記載のディスクアレイ装置。

【請求項 4】 前記制御部を、物理的に独立させて複数個備えたことを特徴とする請求項 1, 2 又は 3 記載のディスクアレイ装置。

【請求項 5】 前記キャッシュメモリは、不揮発メモリであることを特徴とする請求項 1, 2, 3 又は 4 記載のディスクアレイ装置。

【請求項 6】 前記制御部は、前記いずれかのディスクに障害が生じてても当該障害ディスクを縮退せずにデータ読み書き処理を行う、ことを特徴とする請求

項 1, 2, 3, 4 又は 5 記載のディスクアレイ装置。

【請求項 7】 上位ホストからの指令により複数のディスクに対してデータを読み書きするディスクアレイ装置におけるデータ書き込み方法において、

上位ホストにて用いられる論理アドレスに関連付けられたデータを、前記ディスクに対してデータの書き込み処理を行う前に物理アドレスに関連付けて一時的にキャッシュメモリに格納し、

前記キャッシュメモリ上の前記物理アドレスに関連付けられたデータを、当該物理アドレスに対応する前記ディスク上のデータに対して優先して書き込み処理する、ことを特徴とするディスクアレイ装置を用いたデータ書き込み方法。

【発明の詳細な説明】

【0 0 0 1】

【発明の属する技術分野】

本発明は、ディスクアレイ装置にかかり、特に、上位ホストからの指令にて複数のディスクに対してデータを読み書きするディスクアレイ装置に関する。

【0 0 0 2】

【従来の技術】

ディスクアレイ装置では、複数のディスクをグループ化して、データに冗長性を持たせて格納するので、単一のディスク障害によってもデータが損失せず、データ処理を継続することができる。そして、データに冗長性を持たせる方法としては複数あり、R A I D レベルと呼ばれる。複数の R A I D レベルのなかで、R A I D 5 は容量効率に優れているため、R A I D 1 と並んでとくに有用で普及している。R A I D レベルについては、1 9 8 7 年カリフォルニア大学バークレイ校において論文化された、デビット・A・パターソン、ガスギブソン、ランディ・カッツ教授による「A Case for Redundant Arrays of Inexpensive Disks」に詳細に解説されている。また、実用されているディスクアレイ装置の例として、特許文献 1 に開示されている。

【0 0 0 3】

R A I D 技術によれば単一のディスク障害によってデータが損失することは無いが、R A I D 制御を行うディスクアレイ装置内の制御部であるディレクタの障

害に関してはRAID技術の範囲外である。従って、ディスク障害によってデータ損失がないので、ディレクタの障害によってもデータ損失せず、データ処理の継続が行われる装置が望ましい。そのために、ディレクタを二重化し、単一のディレクタ障害が発生しても、別のディレクタによって処理を継続できるようにするのが一般的である。但し、RAID5においてディレクタ障害が発生したときに、データコヒーレンシ、すなわち、ディスクとメモリとのデータの同一性において問題が生じるので、それを図8を参照して説明する。

#### 【0004】

図8(a)、(b)、(c)を用いて、RAID5における書き込み処理(ライト処理)を説明する。ディスク101~105でRAID5を構成していて、データを記憶する領域(ストライプ)111~115が形成されている。そして、領域111~114にはユーザーデータが格納されていて、領域115には領域111~114のパリティ情報が格納されている。

#### 【0005】

ここで、領域111に対してデータ121を書き込む(ライトする)場合について説明する。ライトする場合、領域111内を新しいデータに更新するだけでなく、領域115も新しいデータに対応したパリティに更新しなくてはならない。従って、まず、図8(a)で示すように、ライト動作に先立って、領域111と領域115から、旧データ122と旧パリティ123を読み出す。次に、図8(b)で示すように、書き込みデータ121、旧データ122、旧パリティ123の3つのデータから新パリティ124を生成する。このようにパリティを生成するときには並列処理を可能とするためにディスク112~114のアクセスを行わない方法で行う。最後に、図8(c)で示すように、書き込みデータ121と新パリティ124を、それぞれディスク101、105に書き込む。

#### 【0006】

##### 【特許文献1】

特開2001-344076号公報

#### 【0007】

【発明が解決しようとする課題】

以上の処理過程において、障害が発生したときの回復処理を以下に説明する。  
図 8 (c) において、書き込みデータが記憶されるデータディスク 101 とパリティデータが記憶されるパリティディスク 105 へのライト時に、いずれか一方には書き込むことができたが、他方には書き込めなかった場合に、単純に図 8 (a) ~ (c) の処理をすべて最初からやり直すとする。すると、パリティが不正な値になってしまうという不都合が生じる。パリティが不正な値になると、ディスクのいずれかが縮退した場合に、不正なパリティを用いてデータ普及することになり、データ化けとなってしまう、データ読み書きの信頼性の低下という問題が生じる。

#### 【0008】

そして、図 8 (c) において一方のディスクに対しては書き込むことができたが、他方のディスクには書き込めなかった場合には、書き込めなかったディスクを縮退しなければならず、処理が遅延したり、ディスクの交換による運用コストが増大するという問題も生じる。

#### 【0009】

また、書き込み処理中にディレクタ（制御部）障害が発生したために書き込めなかった場合に、他のディレクタである代替ディレクタが書き込みデータ 121 を見つけて、図 8 (a) からの処理を行うことも考えられる。しかし、かかる場合には、上述したようにパリティ不正が発生してしまい、信頼性の低下という問題が生じる。

#### 【0010】

##### 【発明の目的】

本発明は、上記従来例の有する不都合を改善し、特に、ディスクに障害が生じた場合であっても、データコヒーレンスを維持し、信頼性の高いディスクアレイ装置を提供することをその目的とする。

#### 【0011】

##### 【課題を解決するための手段】

そこで、本発明では、上位ホストからの指令により複数のディスクに対してデータを読み書き制御する制御部と、ディスクに対して読み書きするデータを一時

的に記憶するキャッシュメモリとを備え、制御部が、キャッシュメモリ上において、上位ホストにて用いられる論理アドレスに関連付けたデータを物理アドレスに関連付けて前記ディスクに対して読み書き制御を行うディスクアレイ装置において、制御部が、ディスクに対して読み書き制御を行う際に、キャッシュメモリ上の物理アドレスに関連付けられたデータを当該物理アドレスに対応するディスク上のデータに対して優先して処理する、という構成を採っている。

#### 【0012】

このような構成にすることにより、ディスクに対して書き込み、読み出し処理を行っている最中に、ディスク障害、あるいは、制御部に障害が発生し、ディスク上のデータが不定な状態になっても、物理アドレスに関連付けられたデータを用いて読み書き処理を継続することで、データの安定性、具体的には、データコヒーレンシの維持を図ることができ、データの信頼性の向上を図ることができる。

#### 【0013】

また、制御部が、ディスクに対してデータの書き込み処理を行う前に、当該ディスクに書き込むデータを物理アドレスに関連付けてキャッシュメモリに格納する。

#### 【0014】

これにより、ディスクに書き込まれるデータは、書き込まれる前に物理アドレスに関連付けられてキャッシュメモリに記憶されるため、かかる状態で制御部に障害等が発生した場合であっても必ずキャッシュメモリに残ることとなる。従って、その後、当該キャッシュメモリ上の物理アドレスに関連付けられたデータがディスク上のデータに優先されて当該物理アドレスを参照してディスクに書き込まれるため、障害前と同様の書き込み処理を継続でき、データコヒーレンシの維持を図ることができる。

#### 【0015】

また、制御部が、ディスクにデータ書き込み処理を行うと共に当該書き込みが完了したことを確認した後に、キャッシュメモリ上で物理アドレスに関連付けられた書き込みデータを当該物理アドレスに関連付けられた状態から解除する。



**【0016】**

これにより、確実に書き込み処理が完了したことを確認した後にキャッシュメモリ上にて物理アドレスに関連付けられた状態から解除されるため、完全にデータ書き込み処理が終了しない限りはキャッシュメモリに物理アドレスに関連付けられた書き込みデータが残ることとなる。従って、上述したように、当該データが後に優先して読み書き処理されるため、障害前の処理を継続でき、より信頼性の向上を図ることができる。

**【0017】**

また、制御部を、物理的に独立させて複数個備えると望ましい。これにより、一つの制御部に障害が生じたとしても、別の制御部がキャッシュメモリ内の物理アドレスに関連付けられたデータの優先処理を引き継ぐことで、データコヒーレンシの維持を図ることができる。

**【0018】**

また、キャッシュメモリは、不揮発メモリであると、障害によってディスクアレイ装置自体の動作が停止したとしても、キャッシュメモリには物理アドレスに関連付けられたデータが残っており、かかるデータに対して処理を継続することで、データコヒーレンシの維持を図ることができる。

**【0019】**

さらに、制御部は、いずれかのディスクに障害が生じても当該障害ディスクを縮退せずにデータ読み書き処理を行う。これにより、障害ディスクをすぐに縮退することなくデータ処理を継続するため、発生した障害が一時的なもの、あるいは、局所的なものなどの軽障害である場合には、ディスク交換する必要がないため、運用コストの削減を図ることができる。

**【0020】**

また、本発明では、上位ホストからの指令により複数のディスクに対してデータを読み書きするディスクアレイ装置におけるデータ書き込み方法であって、上位ホストにて用いられる論理アドレスに関連付けられたデータを、ディスクに対してデータの書き込み処理を行う前に物理アドレスに関連付けて一時的にキャッシュメモリに格納し、キャッシュメモリ上の物理アドレスに関連付けられたデー

タを、当該物理アドレスに対応するディスク上のデータに対して優先して書き込み処理する、というディスクアレイ装置を用いたデータ書き込み方法をも提供している。

#### 【0021】

このようにしても、上述と同様の作用・効果を発揮し、上記目的を達成することができる。

#### 【0022】

##### 【発明の実施の形態】

以下、本発明の一実施形態を、図1乃至図7を参照して説明する。図1は、本発明におけるデータ処理の概略を説明する説明図である。図2は、本発明の構成を示すブロック図であり、図3は、キャッシュメモリ内におけるデータ構成を示すブロック図である。図4乃至図7は、データ処理の動作を示すフローチャートである。

#### 【0023】

本発明におけるディスクアレイ装置は、パーソナルコンピュータやサーバコンピュータなどの上位ホストからの指令により、RAID5によって複数のディスクに対してデータを読み書きするものである。このとき、ディスクアレイ装置は、データの読み書き処理を制御部であるディレクタにて制御し、また、ディスクに対して読み書きするデータをキャッシュメモリに一時的に記憶する。そして、キャッシュメモリ上においては、ディレクタが、上位ホストにて用いられる論理アドレスに関連付けたデータを物理アドレスに関連付け、そして、ディスクに対して読み書き制御を行っている。

#### 【0024】

まず、図1を参照して、上述したようなディスクアレイ装置における本発明の特徴を説明する。

#### 【0025】

図1では、ディスク11～15によりRAID5を構成している。ここで、RAID5とは、RAID技術の1つであり、データをディスクに記録する際に、複数のディスクにデータを分散して書き込むと同時に、パリティを計算及び生成

してディスクに書き込む。そして、パリティ用ディスクは特に決まっておらず、全ディスク分散して書き込む、というものである。そして、従来の技術において説明したように、データに冗長性を持たせたディスクへの書き込み方式の一つである。

#### 【0026】

また、ディスクに対して読み書きされるデータが一時的に記憶されるキャッシュメモリ 30 上には、データを格納するための領域であるキャッシュページ（例えば、ライトデータ、新パリティデータなどが格納されている領域）が存在する。そして、キャッシュページは、論理ドメイン 31、物理ドメイン 32、ワークドメイン 33 と名付けられたいずれかの領域に属している。ここで、論理ドメイン 31 とは、論理アドレスに関連づけられたデータの属する場所であり、物理ドメインとは、物理アドレス 32 に関連づけられたデータの属する場所である。また、ワークドメイン 33 とは論理アドレスにも物理アドレスにも関連づけられていないデータの属する場所である。但し、後述するように、実際には図 1 のように各ドメイン毎に領域が分けられているわけではない。説明の便宜上、図 1 のように示したまでである。

#### 【0027】

今、キャッシュメモリ 30 上に上位ホスト（図示せず）からのライトデータ 41 が存在し、これをディスクに書き込むが、このデータは論理アドレスで検索可能なキャッシュページに格納されているので、論理ドメイン 31 に属している。図 1 では、このデータをライトするのはディスク 11 であり、対応するパリティはディスク 15 に存在するので、ディスク 11 と 15 において上記論理アドレスに対応するアドレスから、あらかじめ旧データ 43 と旧パリティ 44 を読み出す（矢印 A1, A2 参照）。具体的には、旧データ 41 は、ライトデータ 41 が関連付けられている論理アドレスに対応するディスク 11 上の領域（アドレス）21 に格納されているデータであり、同様に、旧パリティ 44 は、ディスク 15 の領域 25 に格納されているデータであって、各ディスク 11, 15 の領域 21, 25 から読み出す。ちなみに、他のディスク 12～14 にも、他のデータに対応する領域 22～24 がそれぞれ形成されている。

## 【0028】

そして、論理ドメイン 31 内のライトデータ 41 と、ワークドメイン 33 内の旧データ 43、旧パリティ 44 とから、新パリティ 45 をワークドメイン 33 内に生成する（矢印 A3、A4、A5 参照）。ここで、旧データ 43、旧パリティ 44、新パリティ 45 がワークドメイン 33 のキャッシュページに属しているのは、読み書き処理のための一時的なデータであるからである。

## 【0029】

次に、ライトデータ 41 と新パリティ 45 をディスク 11、15 に書き込む処理を行うが、本発明においてはディスクへの書き込み処理を行う前に、ライトデータ 41 と新パリティ 45 とをドメイン変換により、物理ドメイン 32 のキャッシュページとする。すなわち、ライトデータ 41 は、そもそも上位ホスト（図示せず）からの指令によりディスクアレイ装置に送られてきたため、論理ドメイン 31 内では論理アドレスに関連付けられて管理されており、これを、物理アドレスに関連付けられるよう変換を行う（矢印 A6、A7 参照）。

## 【0030】

その後、ディレクタは、物理ドメイン 32 内のキャッシュページを、ディスク上の該当アドレスデータよりも優先して、書き込み処理を行う（矢印 A8、A9 参照）。すなわち、ディスクへの書き込み処理が行われる前か、実施中か、完了しているかに関わらず、ディスク上の該当アドレスのデータに対して、物理ドメインのキャッシュページが優先される。そのため、ディスク書き込み中にディレクタ障害が発生して、ディスク上のデータが不定な状態になっても、物理ドメイン 32 のキャッシュページにライトデータが残っているため、かかるデータが再度書き込まれることにより、データコヒーレンシが維持されることになる。

## 【0031】

次に、図 2 乃至図 7 を参照して、本発明の具体的な実施例を説明する。まず、図 2 に示すように、本発明であるディスクアレイ装置 50 は、データの読み書きを制御する制御部であるディレクタ 51、52 を 2 つ備えている。このディレクタ 51、52 は、SCSI などの汎用インターフェースによって上位ホストであるホストコンピュータ 60 に接続され、当該ホストコンピュータ 60 から受領し

たコマンドを処理する。またディレクタ 51, 52 は、やはり汎用インターフェースによってディスク 54 ~ 59 に接続され、ホストコンピュータ 60 から転送されたデータを、ディスクの適当な場所に格納したり、必要なデータを読み出したりする。ここで、ディレクタ 51, 52 は、2 つ備えてられていることを例示したが、必ずしもこの個数に限定されない。1 つでもよく、3 つ以上であってもよい。また、ディレクタ 51, 52 は、それぞれ物理的に独立して形成されている。すなわち、一つの CPU 内に 2 つの機能が存在するよう構成されているのではなく、図 2 の例では、2 つの個別のハードウェアにて構成されている。

#### 【0032】

さらに、ディレクタ 51, 52 は、同一の記憶手段である共有メモリ 53 に接続されているが、この共有メモリ 53 は、キャッシュメモリとして使用され、不揮発なメモリでもある。そして、ディレクタ 51, 52 は、ホストコンピュータ 60 とやりとりするデータを、共有メモリ 53 に一旦格納することでホストからコマンドに高速に応答することができる。なお、共有メモリ 53 は、不揮発なメモリにて構成されていなくてもよい。

#### 【0033】

次に、上記キャッシュメモリとして機能する共有メモリ 53 内のデータ構造について、図 3 を参照して説明する。共有メモリ 53 上には、論理ドメイン検索エントリテーブル 71、物理ドメイン検索エントリテーブル 72、キャッシュページ配列 80 が存在する。そして、論理ドメイン検索エントリテーブル 71 には、論理アドレスから一意に決まるポインタ 71a ~ 71d であって、その参照先に当該論理アドレスに関連づけられるキャッシュページがある。従って、論理ドメイン検索エントリテーブル 71 にて、いずれかのポインタ 71a ~ 71d から、論理ドメイン 31 に属するキャッシュページを検索することができる。同様に、物理ドメイン検索エントリテーブル 72 には、物理アドレスから一意に決まるポインタ 72a ~ 72d のいずれかのポインタから、当該物理アドレスに関連づけられるキャッシュページを検索することができる。すなわち、検索されたキャッシュページは、物理ドメイン 32 に属しているキャッシュページである。

#### 【0034】

また、キャッシュページ配列 80 には、複数のキャッシュページ 81～91 と、各キャッシュページに対応する未書き込みフラグ 81 f から 91 f の領域とがある。そして、キャッシュページには、ディスクに対して読み書きされるデータ（ライトデータやパリティデータなど）が格納される。また、すべてのキャッシュページ 81～91 は、上述した論理ドメイン 31、物理ドメイン 32、ワークドメイン 33 のいずれかに属する。

#### 【0035】

ここで、図 3 の矢印に示すように、キャッシュページ 81, 82, 83, 87 は論理ドメイン検索エントリテーブル 71 から検索でき、従って、これらキャッシュページは、論理ドメイン 31 に属している。また、物理ドメイン 32 のキャッシュページ 84, 91 は物理ドメイン検索エントリテーブル 72 から検索できるようになっている。そして、残りのキャッシュページ、すなわち、論理アドレスにも物理アドレスにも関連づけられていないキャッシュページ 85, 86, 88, 89, 90, 91 は、ワークドメイン 33 に属するキャッシュページである。

#### 【0036】

また、図 2 に示すディレクタ 51, 52 は、以下に説明する機能を有する。まず、ディスクに対して読み書き制御を行う際に、共有メモリ（キャッシュメモリ）53 上の物理アドレスに関連付けられたデータを当該物理アドレスに対応するディスク上のデータに対して優先して処理する機能を有する。従って、物理ドメイン 32 に属するデータがある場合には、このデータに対する処理が、ディスクから読み出し処理や当該ディスクに対する他のデータの書き込み処理などよりも優先して行われる。

#### 【0037】

さらに、ディレクタ 51, 52 には、ディスクに対する書き込み処理を行う前に、ディスクに書き込むデータを物理アドレスに関連付けて共有メモリ（キャッシュメモリ）53 に格納する機能を有する。従って、書き込み処理の対象となっているデータは、書き込み処理前に常に物理ドメイン 32 内に格納されることとなる。そして、ディレクタ 51, 52 は、ディスクにデータ書き込み処理を行な

った後に、当該書き込みが完了したことを確認する機能を有し、この機能は、当該書き込み完了を確認した後に、当該書き込みデータをキャッシュメモリ上で物理アドレスに関連付けられた状態から開放する。すなわち、物理ドメイン 32 から移動あるいは削除される。従って、書き込み対象のデータは、完全にディスクに書き込まれない限りは物理ドメイン 32 に残ることとなり、その後、当該物理ドメイン 32 内のデータが、ディスク上のデータよりも優先してディレクタ 51, 52 にて処理される。

#### 【0038】

また、ディレクタ 51, 52 には、いずれかのディスクに障害が生じても当該障害ディスクを縮退せずにデータ読み書き処理を行う機能を有する。すなわち、軽障害が発生した程度では読み書き処理を停止せずに読み書き処理を続行する。

#### 【0039】

そして、本実施形態では、ディレクタ 51, 52 を 2 つ備えているが、このように複数のディレクタが備えられている構成においては、各ディレクタが他のディレクタの状況を監視し、当該他のディレクタに障害が生じたら、障害が生じたディレクタの処理を引き継いでディスクへの読み書き処理を行う。例えば、ディスクに共有メモリ 53 内の物理ドメイン 32 に格納されているデータを書き込む際に、一方のディレクタ 51 に障害が生じたら、他方のディレクタ 52 は当該物理ドメイン 32 のデータを優先的に処理し、障害前と同様にディスクへの書き込み処理を継続して実行する。

#### 【0040】

ここで、ディレクタ 51, 52 は、上述した機能の全てを必ずしも備えていることに限定されない。そのうちの一部の機能が備わっていてもよい。また、上記機能は、あらかじめ各機能用プログラムが CPU であるディレクタ 51, 52 に組み込まれており、あるいは、不揮発メモリなどの記憶手段に記憶されてこれを読み出すことにより、ディレクタ 51, 52 内に各機能が構築され、これにより実現できる。なお、上記機能については、次の動作説明時に詳述する。

#### 【0041】

次に、図 4 乃至図 7 のフローチャートを参照して、本実施形態におけるディス

クアレイ装置 50 の動作を説明する。

**【0042】**

まず、図 4 のリード処理について説明する。はじめに、ディスクアレイ装置 50 のディレクタ 51、52 が、ホストコンピュータ 60 からディスク上の所定のデータを読み出すようリードコマンドを受信する（ステップ S1）と、共有メモリ 53 内の論理ドメイン検索エントリテーブル 71 を用いて、リード対象となる論理アドレスにデータがあるか、すなわち、論理ドメインキャッシュページがあるか否かを調べる（ステップ S2）。以下、キャッシュページがあるか否かの判定をヒット判定と呼び、キャッシュページがあることをヒットするという。

**【0043】**

そして、キャッシュページがヒットした場合、すなわち、対応するキャッシュページがある場合には（ステップ S2 で肯定判断）、当該キャッシュページからホストコンピュータ 60 にデータ転送を行う（ステップ S8）。逆に、キャッシュページがヒットしなかった場合には、論理アドレスからそれに対応する物理アドレスを算出してアドレス変換する（ステップ S3）。そして、物理ドメイン検索エントリテーブル 72 を用いて、変換した物理アドレスにデータがあるか、すなわち、物理ドメインキャッシュページのヒット判定を行う（ステップ S4）。

**【0044】**

ここで、キャッシュページがヒットした場合には（ステップ S4 にて肯定判断）、当該キャッシュページからワークドメイン 33 のキャッシュページにデータコピーを行う（ステップ S5）。また、キャッシュページがヒットしなかった場合には（ステップ S4 にて否定判断）、ディスクからワークドメイン 33 のキャッシュページにデータをコピーする（ステップ S6）。すると、いずれの場合にも、ワークドメイン 33 のキャッシュページに必要なデータが格納されるので、データを格納したワークドメイン 33 のキャッシュページを論理ドメイン 31 のキャッシュページにドメイン変換する（ステップ S7）。この処理は、具体的には、ワークドメイン 33 のキャッシュページに、論理ドメイン検索エントリテーブル 71 のポインタを参照させるよう、当該ポインタを書き換えることにより行う。これにより、当該キャッシュページを論理アドレスから検索できるようにな



る。その後、論理ドメインキャッシュページからホストコンピュータ60にデータを転送する（ステップS8）。以上の処理によってリードコマンド処理は完了となる。

#### 【0045】

次に、図5のライト動作を説明する。まず、ディレクタ51、52が、ホストコンピュータ60からデータをディスクに記録するというライトコマンドを受信すると（ステップS11）、論理ドメイン検索エントリテーブル71を用いて、その論理アドレスに対応する論理ドメインキャッシュページがあるか否かを判定する（ヒット判定、ステップS12）。そして、キャッシュページがヒットした場合には（ステップS12で肯定判断）、ホストコンピュータ60からそのキャッシュページにライトデータを転送する（ステップS14）。このとき、当該キャッシュページに付随する未書き込みフラグをセットする。

#### 【0046】

また、ステップS12にてキャッシュページがヒットしなかった場合には、ホストコンピュータ60からライトデータをワークドメイン33のキャッシュページにデータ転送する（ステップS13）。そして、データを格納したワークドメイン33のキャッシュページを、論理ドメイン31のキャッシュページにドメイン変換する（ステップS15）。以上の処理によってライトコマンド処理は完了となる。

#### 【0047】

続いて、上記ライトコマンド処理によってキャッシュメモリに格納されたデータをディスクに書き込む処理を、図6のフローチャートを参照して説明する。ここで、ディレクタ51、52では、上述したコマンド処理動作とは非同期に、論理ドメイン31の未書き込みデータの監視処理が、定期的に行われる（ステップS21）。そして、監視処理は、具体的には、論理ドメイン31の未書き込みフラグがセットされたキャッシュページを検索することにより行われる（ステップS22）。

#### 【0048】

このとき、未書き込みフラグがセットされたキャッシュページが存在する場合

には（ステップS 2 2で肯定判断）、そのキャッシュページの論理アドレスからそれに対応する物理アドレスを算出し、すなわち、アドレス変換し（ステップS 2 3）、その物理アドレスにおいて物理ドメイン検索エントリテーブル7 2を用いて、物理ドメイン3 2のキャッシュページのヒット判定を行う（ステップS 2 4）。

#### 【0049】

そして、キャッシュページがヒットした場合には、ライトデータは既に物理アドレスに関連付けられているため、このときには当該キャッシュページの書き込み処理は実行せず（ステップS 2 4にて肯定判断）、後の処理において書き込む（図7参照）。そして、物理ドメインキャッシュページがヒットしなかった場合には（ステップS 2 4にて否定判断）、ワークドメイン3 3のキャッシュページに、該当するディスクから旧データと旧パリティデータを読み出す（ステップS 2 5、図1の符号4 3、4 4参照）。そして、旧データ、旧パリティ及びライトデータとを用いて、ワークドメイン3 3のキャッシュページに新パリティデータを生成する（ステップS 2 6、図1の符号4 5参照）。

#### 【0050】

続いて、ライトデータと新パリティを物理ドメイン3 2にドメイン変換する（ステップS 2 7）。この処理は、具体的には、論理ドメイン検索エントリテーブル7 1のポインタと物理ドメイン検索エントリテーブル7 2のポインタを書き換えることで、ライトデータと新パリティデータを物理アドレスから検索できるようにする。また、同時に、そのキャッシュページの未書き込みフラグをリセットする。

#### 【0051】

その後、ドメイン変換したキャッシュページからディスクへデータ転送を行い、実際にライト処理を行う（ステップS 2 8）。そして、ディスクにライト処理を行った結果、エラーが発生していないかの判定を行い（エラー判定、ステップS 2 9）、エラーが発生していなければ（ステップS 2 9にて否定判断）、ライトデータ及び新パリティデータを削除する（ステップS 3 0）。この処理は、具体的には、物理ドメイン検索エントリテーブル7 2のポインタを書き換えること

で、当該キャッシュページをアドレスで検索できないようにし、ワークドメイン 3 3 のキャッシュページとする処理である。すなわち、物理アドレスに関連付けられた状態から開放する処理である。一方、ライト処理にエラーがあった場合には（ステップ S 2 9 にて肯定判断）、ライトデータ及び新パリティデータを物理ドメイン 3 2 に残したまま処理を終了する。

#### 【 0 0 5 2 】

次に、ディスクライト処理において残った物理ドメインのキャッシュページをディスクに書き込む処理を、図 7 のフローチャートを用いて説明する。このとき、ディレクタ 5 1, 5 2 では、コマンド処理動作とは非同期に、物理ドメイン 3 2 のキャッシュページを定期的に監視している（ステップ S 3 1）。具体的には、物理ドメイン 3 2 のキャッシュページを検索する（ステップ S 3 2）。

#### 【 0 0 5 3 】

そして、物理ドメイン 3 2 のキャッシュページが存在する場合には（ステップ S 3 2 にて肯定判断）、当該キャッシュページからディスクにデータ転送を行う（ステップ S 3 3）。すなわち、物理ドメインに残ったライトデータ及び新パリティデータを、実際にディスクに書き込む。

#### 【 0 0 5 4 】

その後、ディスクへのライト処理の結果をエラー判定し（ステップ S 3 4）、エラーが発生していなければ（ステップ S 3 4 で否定判断）、当該キャッシュページを削除する（ステップ S 3 5）。この処理は、具体的には上述と同様に、物理ドメイン検索エントリテーブル 7 2 のポインタを書き換えることで、当該キャッシュページをアドレスで検索できなくし、ワークドメイン 3 3 のキャッシュページとする処理である。一方で、エラーがあった場合には（ステップ S 3 4 にて肯定判断）、当該キャッシュページを物理ドメインに残したまま処理を終了する。

#### 【 0 0 5 5 】

そして、上記図 7 に示す物理ドメインの監視処理が常に実行され、物理ドメインに残されているデータ、すなわち、物理アドレスに関連付けられているデータの優先的な書き込み処理が行われる。

**【0056】**

このようにすることにより、ディスクへのライト処理において、書き込むライトデータと、それに伴って更新すべきパリティデータとを、ディスクへのライトを実行する前に、キャッシュメモリ上で物理アドレスによって検索可能な物理ドメインのキャッシュページとして管理することにより、書き込み処理中にディレクタが障害によりダウンした場合であっても、他の代替ディレクタで物理アドレスに関連付けられているデータの優先処理が継続されるため、障害発生前の書き込み処理を継続することができ、データコヒーレンスを維持することができる。その結果、ディスクアレイ装置の信頼性の向上を図ることができる。

**【0057】**

また、ディレクタが二重化されていない場合に当該ディレクタ障害が発生したり、あるいは、二重化されていても電源障害のようにディスクアレイ装置全体が停止してしまうような障害が書き込み処理中に発生した場合であっても、キャッシュメモリを不揮発メモリとすることで、障害回復後の再起動後にも当該不揮発メモリに残されている物理アドレスに関連付けられたデータが優先的に処理されるため、障害発生前の書き込み処理を継続することができ、データコヒーレンスを維持できる。

**【0058】**

さらに、ディスク障害によってエラーが発生したときでも、書き込めなかったデータを物理ドメインのキャッシュページとして管理することで、障害ディスクをすぐに縮退しなくてもデータ処理を継続することができる。そのため、発生したディスク障害が一時的な、または局部的な、軽障害である場合には、そのディスクを使い続けることが可能になり、そのためディスク交換の頻度が下がり、結果的に運用コストを削減することができる。

**【0059】****【発明の効果】**

本発明は、以上のように構成され機能するので、これによると、ディスクに対するデータのリード・ライト処理中に、ディスクや制御部に障害が発生した場合であっても、ディスクに対する処理対象データが物理ドメインのキャッシュペー

ジ上に残り、リード・ライト処理において当該アドレスにアクセスする場合に、そのディスク上のデータよりも物理ドメインのキャッシュページ上のデータが優先されるので、データコヒーレンスを維持したまま処理を継続することができ、リード・ライト処理の信頼性の向上を図ることができる、という従来にない優れた効果を有する。

#### 【0060】

また、ライト処理中に電源障害が発生してディスクアレイ装置全体がダウンした場合でも、キャッシュメモリが不揮発なので書き込み中のデータが物理ドメインのキャッシュページ上に残る。障害が復旧してディスクアレイ装置が再起動するとディレクタはデータコヒーレンスを維持したまま処理を継続することができる。

#### 【0061】

また、物理ドメインのキャッシュページはいずれかのディレクタの定期監視によってディスクに書き込まれ削除されるが、ディスク障害によってデータの書き込みがエラーした場合には書き込めなかったデータが物理ドメインのキャッシュページ上に残るため、ディスク障害が一時的な障害である場合には、当該データが後で定期監視によってディスクに書き込まれ、また、局所的な障害である場合には、物理ドメインのキャッシュページとして残ることになり、ディスクを縮退せずに使い続けることができるため、ディスクの交換コストを削減することができる。

#### 【図面の簡単な説明】

##### 【図1】

本発明の一実施形態におけるデータ処理の概略を説明する説明図である。

##### 【図2】

本発明の一実施形態における構成を示すブロック図である。

##### 【図3】

図2に開示した共有メモリ（キャッシュメモリ）内のデータ構成を示すブロック図である。

##### 【図4】

ディスクアレイ装置によるリード処理の動作を示すフローチャートである。

【図5】

ディスクアレイ装置によるライト処理の動作を示すフローチャートである。

【図6】

図5に示すライト処理によってキャッシュメモリに格納されたデータをディスクに書き込む処理の動作を示すフローチャートである。

【図7】

図6に示すライト処理においてキャッシュメモリの物理ドメインに残ったデータをディスクに書き込む処理の動作を示すフローチャートである。

【図8】

図8（a）～（c）は、従来のディスクアレイ装置におけるデータの書き込み処理を説明する説明図である。

【符号の説明】

- 11～15, 54～59 ディスク
- 21～25 記憶領域（ディスク内）
- 30 キャッシュメモリ
- 31 論理ドメイン
- 32 物理ドメイン
- 33 ワークドメイン
- 41, 42 ライトデータ
- 43 旧データ
- 44 旧パリティデータ
- 45, 46 新パリティデータ
- 50 ディスクアレイ装置
- 51, 52 デイレクタ（制御部）
- 53 共有メモリ（キャッシュメモリ）
- 60 ホストコンピュータ
- 71 論理ドメイン検索エントリテーブル
- 72 物理ドメイン検索エントリテーブル

8 0 キャッシュページ配列

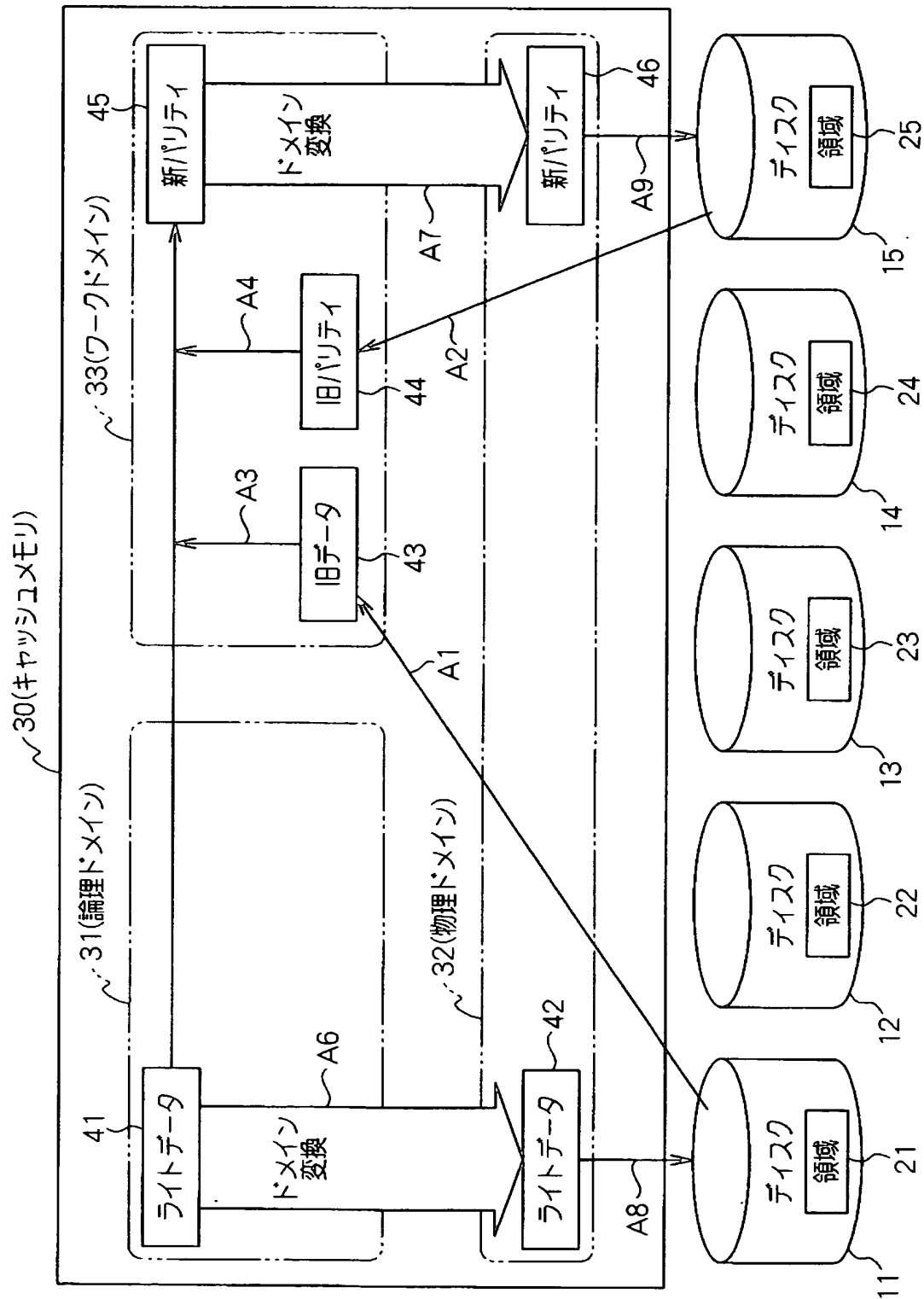
8 1 ~ 9 1 キャッシュページ

8 1 f ~ 9 1 f 未書き込みフラグ

【書類名】

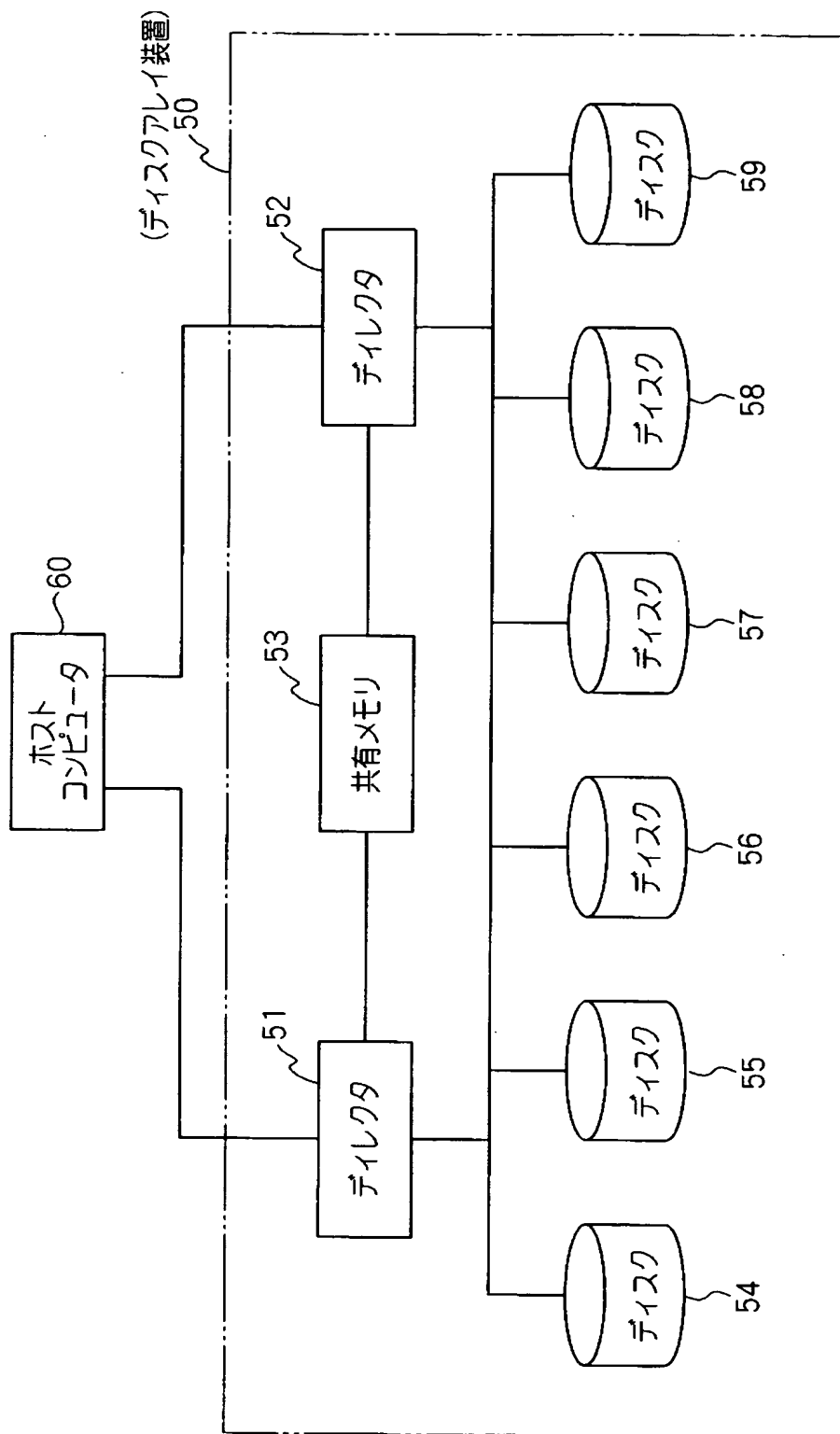
図面

【図1】

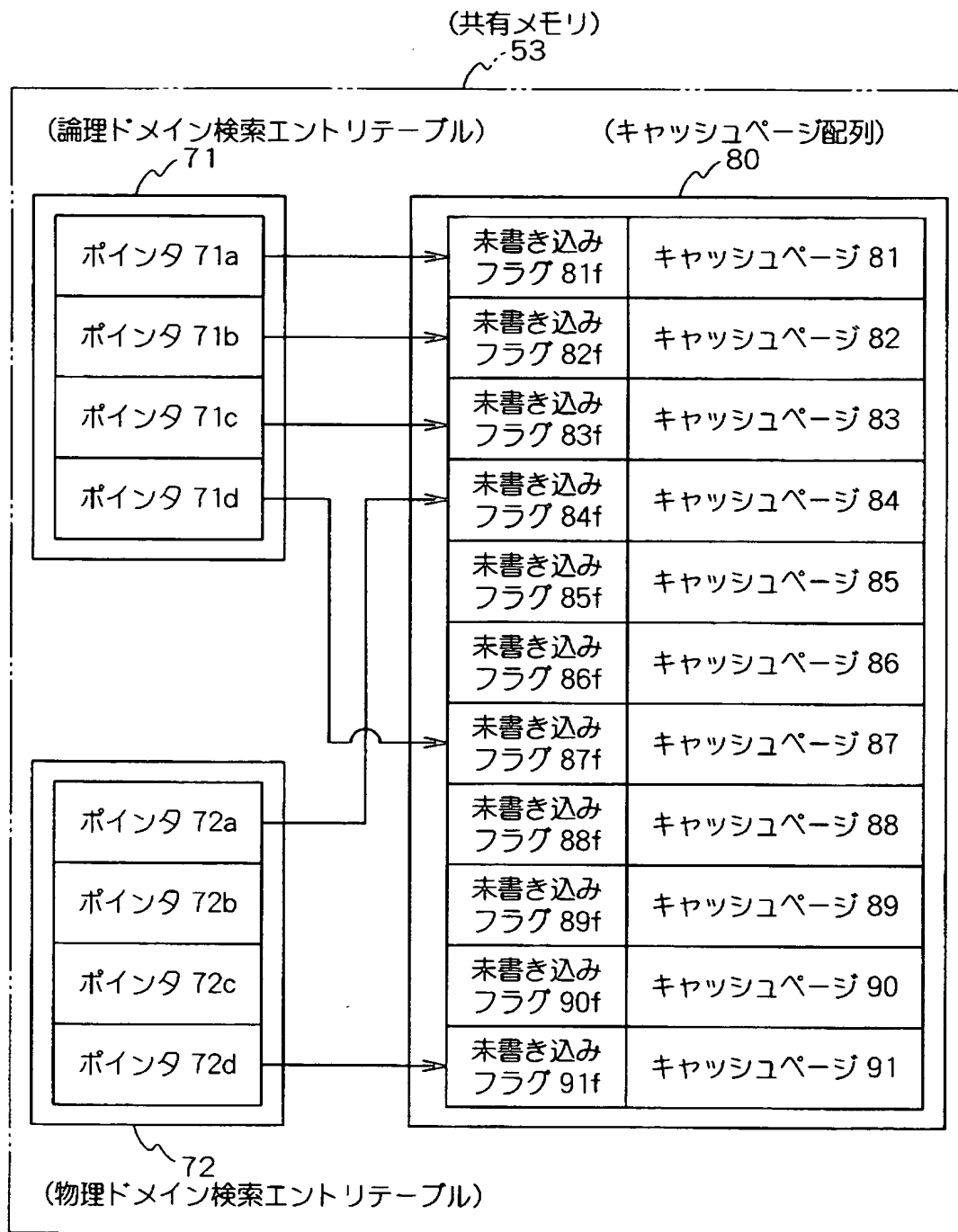




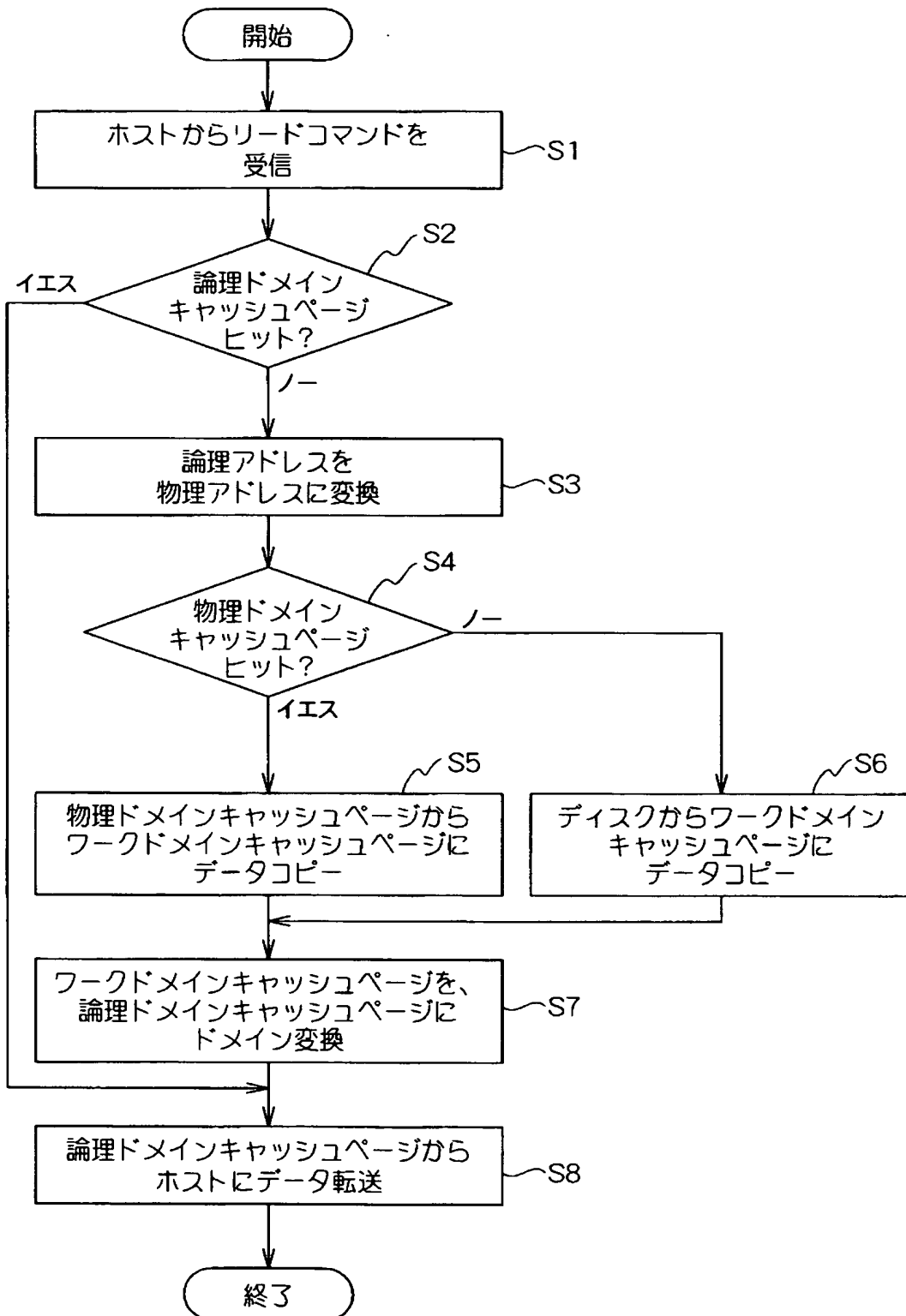
【図 2】



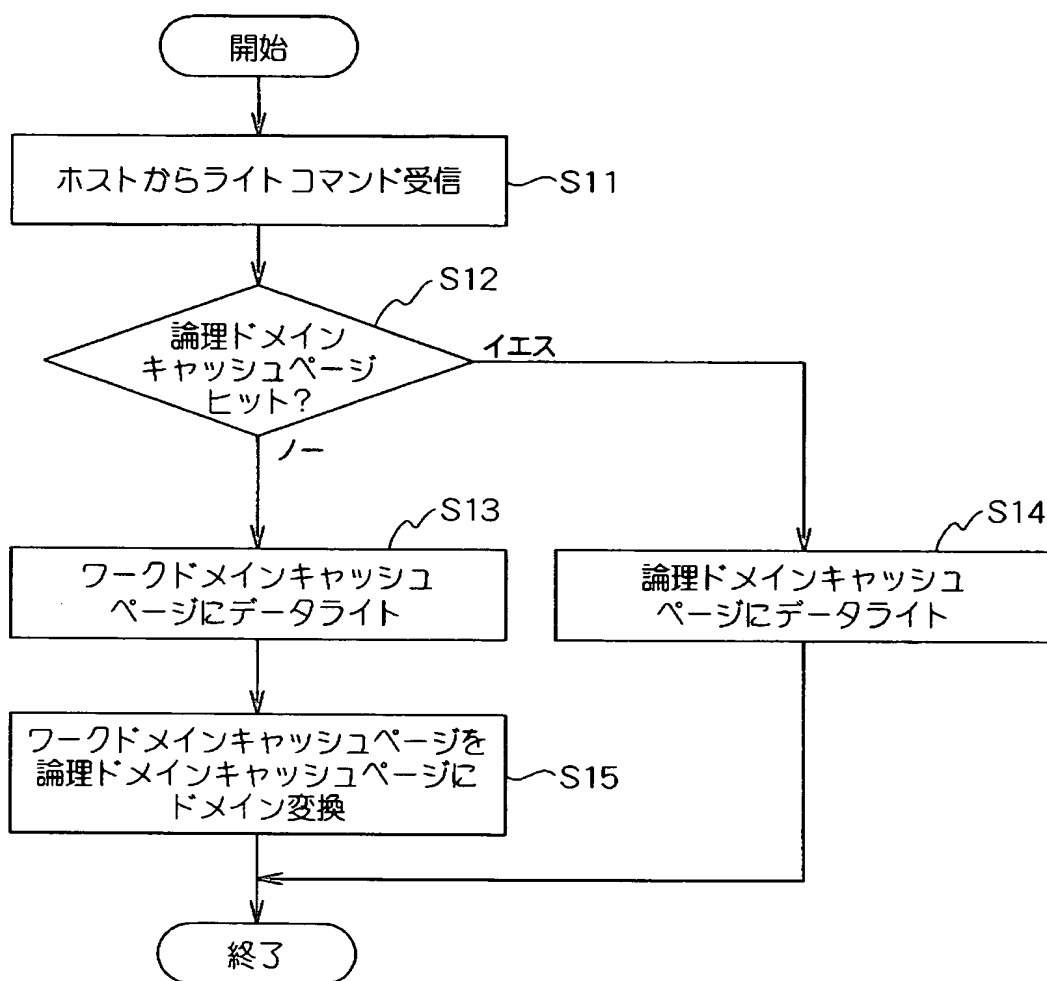
【図 3】



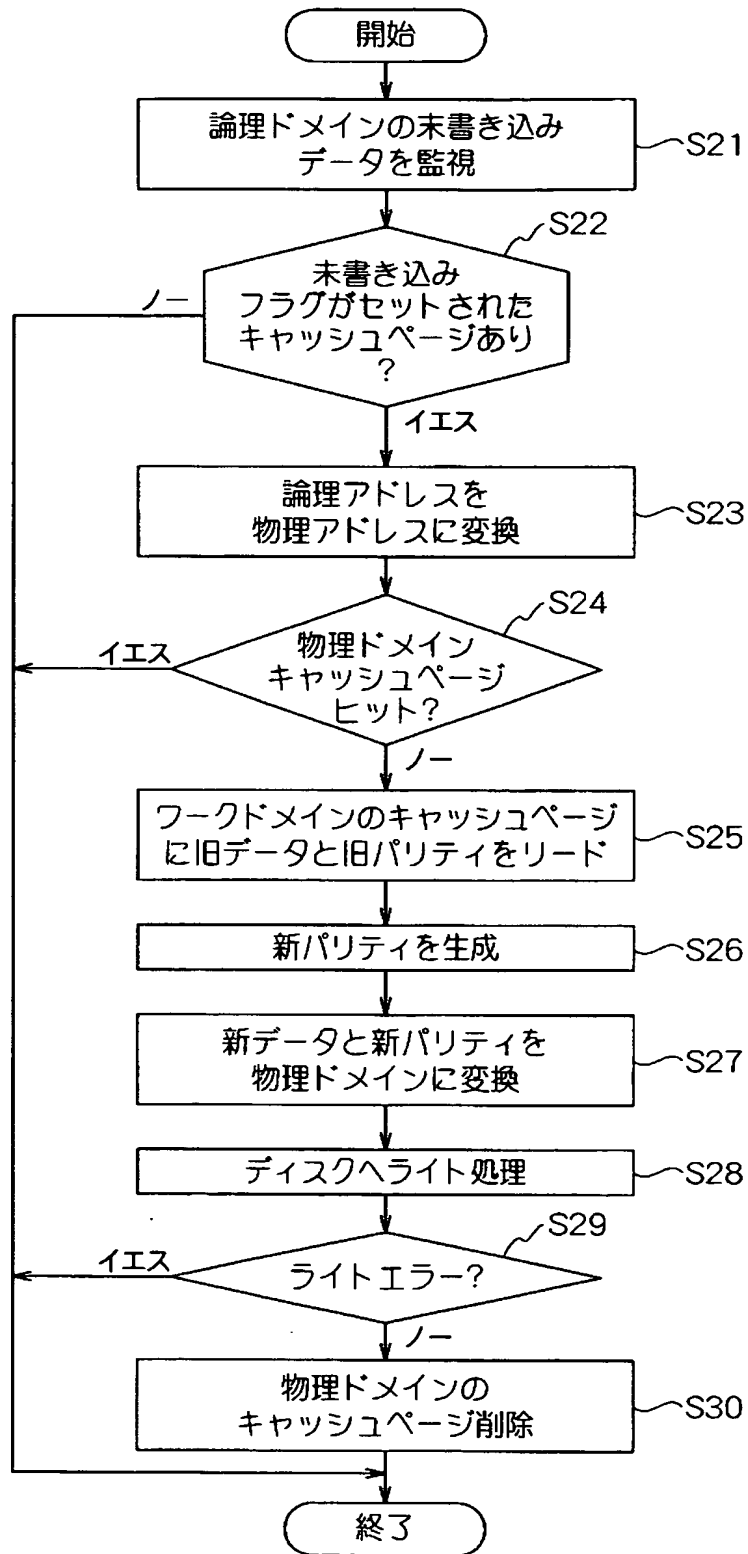
【図 4】



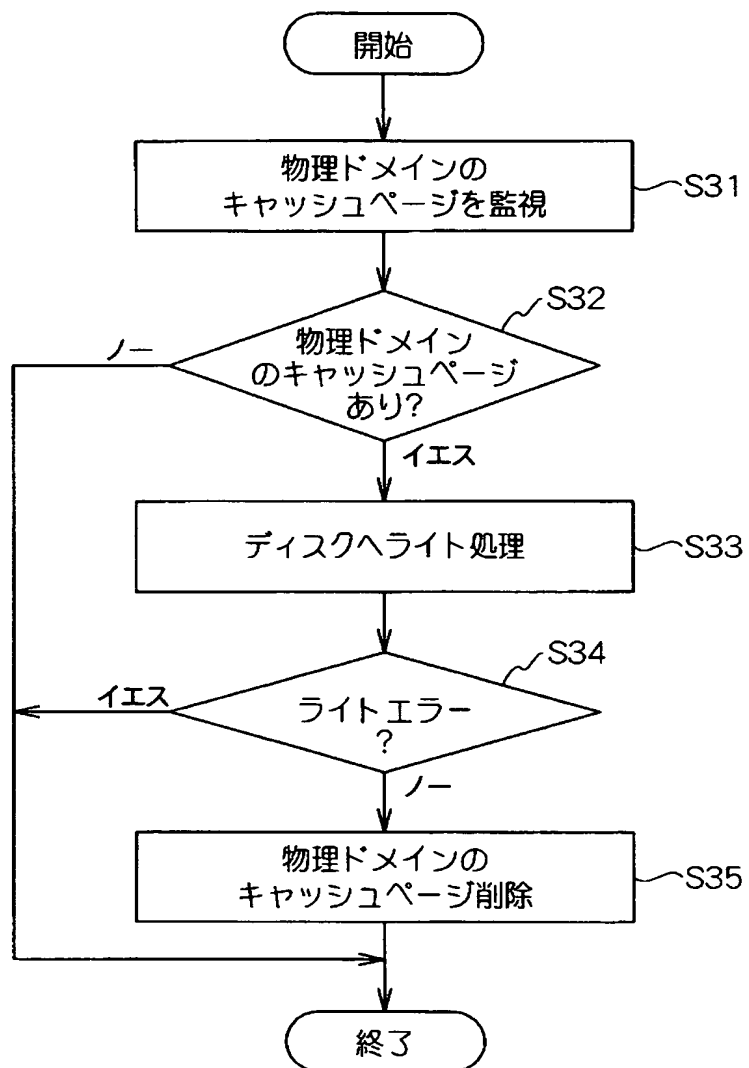
【図 5】



【図 6】

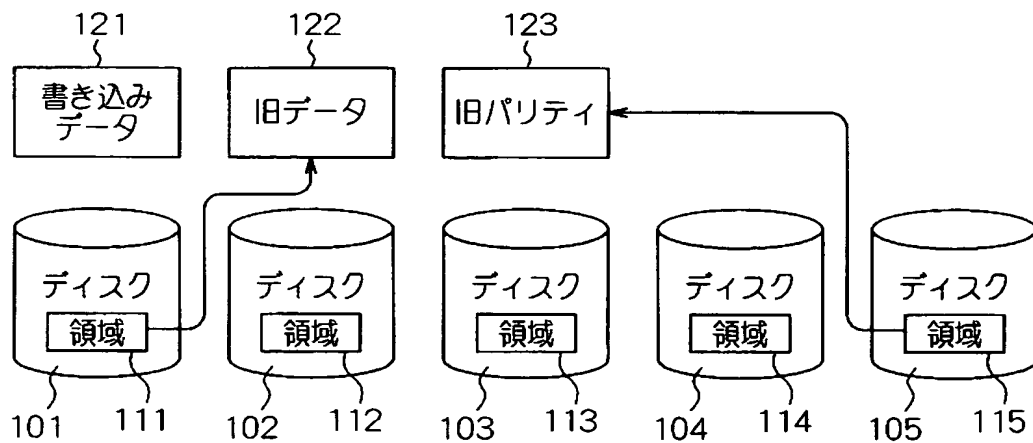


【図 7】

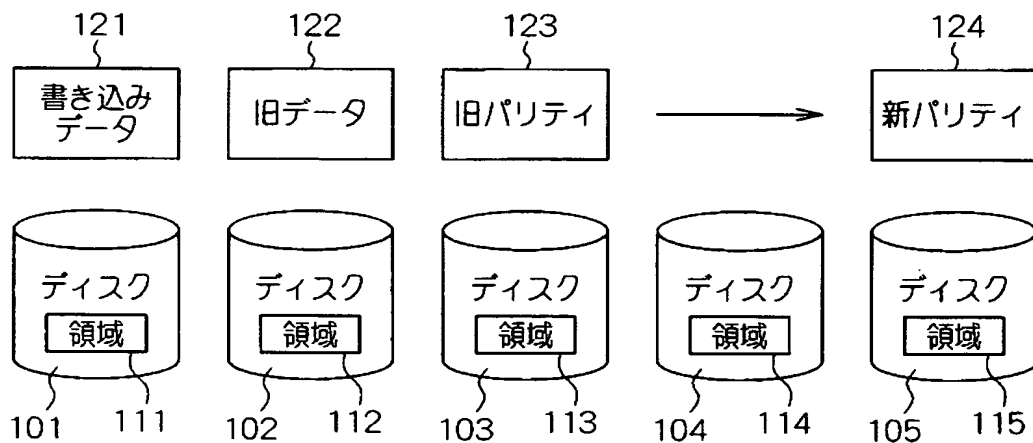


【図 8】

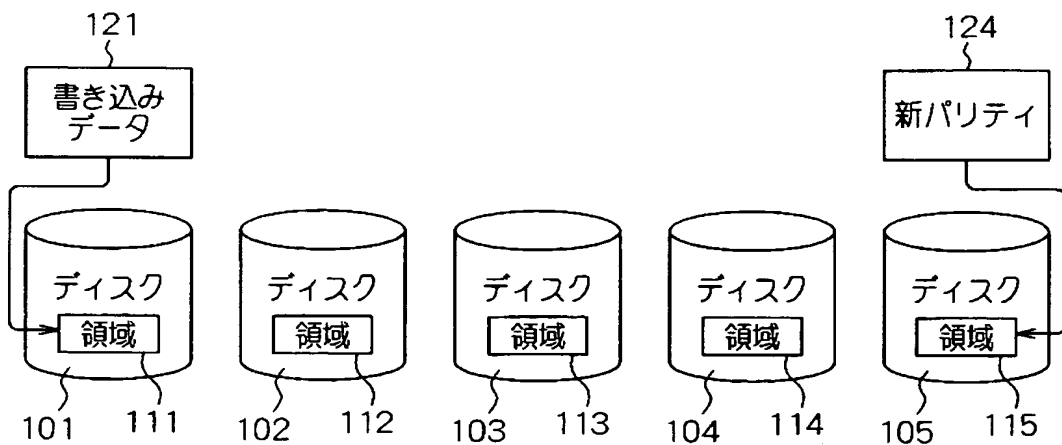
(a)



(b)



(c)



【書類名】 要約書

【要約】

【課題】 ディスクに障害が生じた場合であっても、データコヒーレンスを維持し、信頼性の高いディスクアレイ装置を提供すること。

【解決手段】 上位ホストからの指令により複数のディスクに対してデータを読み書き制御する制御部と、ディスクに対して読み書きするデータを一時的に記憶するキャッシュメモリとを備え、制御部が、キャッシュメモリ上において、上位ホストにて用いられる論理アドレスに関連付けたデータを物理アドレスに関連付けて前記ディスクに対して読み書き制御を行うディスクアレイ装置において、制御部が、ディスクに対して読み書き制御を行う際に、キャッシュメモリ上の物理アドレスに関連付けられたデータを当該物理アドレスに対応するディスク上のデータに対して優先して処理する。

【選択図】 図 1



特願 2 0 0 3 - 0 0 1 3 1 4

出 願 人 履 歴 情 報

識別番号

[ 0 0 0 0 0 4 2 3 7 ]

1. 変更年月日

1 9 9 0 年 8 月 2 9 日

[変更理由]

新規登録

住 所

東京都港区芝五丁目 7 番 1 号

氏 名

日本電気株式会社